J. Appl. Prob. **39,** 630–636 (2002)
Printed in Israel
© *Applied Probability Trust* 2002

POISSON TRAFFIC FLOW IN A GENERAL FEEDBACK QUEUE

EROL A. PEKÖZ * ** AND
NITINDRA JOGLEKAR,* *Boston University*

Abstract

Consider a $\cdot$/G/k finite-buffer queue with a stationary ergodic arrival process and delayed customer feedback, where customers after service may repeatedly return to the back of the queue after an independent general feedback delay whose distribution has a continuous density function. We use coupling methods to show that, under some mild conditions, the feedback flow of customers returning to the back of the queue converges to a Poisson process as the feedback delay distribution is scaled up. This allows for easy waiting-time approximations in the setting of Poisson arrivals, and also gives a new coupling proof of a classic highway traffic result of Breiman (1963). We also consider the case of nonindependent feedback delays.

Keywords: Feedback queue; Poisson convergence; coupling methods

AMS 2000 Subject Classification: Primary 60K25; 90B22
Secondary 60G55; 90B20

1. Introduction

There has been much study of queueing networks with feedback since Jackson (1957) studied networks of M/M/1 queues. More general settings have been studied since the early paper of Takacs (1963), such as general service and exponential feedback times (Foley and Disney (1983)), instantaneous feedback (D'Avignon and Disney (1977/78) and Wortman *et al.* (1991)), and the control of systems with exponential service times (Kumar (1993) and Kuri and Kumar (1997)). But as mentioned by Foley and Disney (1983), waiting times for such systems in general settings are notoriously difficult to analyze. The feedback of customers introduces subtle long-range dependencies so that the resulting stream of customers joining the queue is not even a renewal process.

The assumption of Poisson traffic flow in a queueing network is a common simplifying assumption in modeling practice. There have been many rigorous justifications of this type of phenomenon in different contexts. Melamed (1979) showed that the output process of a network of M/M/1 queues is a Poisson process, and Barbour and Brown (1996) showed that this approximately holds for arcs in such networks provided that the expected number of feedback loops made by customers is small. There are also results in the non-Markovian setting with general arrival and service times; Mountford and Prabhakar (1995) show that the output process of a long sequence of tandem single-server exponential queues converges to a Poisson process, and Prabhakar *et al.* (1996) show that the same holds for sequences of tandem general infinite-server queues. The purpose of the present work is to develop a justification for the approximation of Poisson arrivals in a queueing model with delayed feedback, and thus provide motivation for simple approximations for waiting times.

Received 16 November 2000; revision received 27 May 2002.
* Postal address: School of Management, Boston University, 595 Commonwealth Ave, Boston, MA 02215, USA.
** Email address: pekoz@bu.edu

Our feedback queueing model can be described as a $\cdot$/G/k finite-buffer queue with a rate λ stationary ergodic arrival stream, a general service distribution, and delayed customer feedback. Customers after service may repeatedly return to the end of the queue after a random i.i.d. generally distributed feedback delay distributed as cX for some random variable X with continuous density function and a scaling constant $c > 0$. We refer to c as the 'feedback delay scale factor'; large values of c correspond to a system with large-scale delays. We are interested in the behavior of the system for large values of c. As a stability assumption, we assume that the queue has a finite-sized buffer in the sense that customers arriving to find the total work in the queue above the level b are considered lost for that single pass only (but may still feed back after a delay).

When the number of times customers feed back has finite mean m and finite variance σ^2, we show that the stationary point process of customers returning to the back of the queue converges in distribution to a Poisson process with rate λm as the delay distribution increases in scale by sending $c \to \infty$. If we look at the special case of a Poisson arrival stream, this motivates the approximation of the stationary queue length by that of a no-feedback M/G/k queue with Poisson rate $\lambda(m + 1)$ arrivals. The main result here is perhaps surprising since the Poisson process turns up even with such general interarrival, service, and feedback times. Similar results hold in more general cases where the feedback delays are large in scale but not i.i.d. (see Remark 3.1 below).

Our interest in this queueing system arises from its applicability to modeling software product development processes, and to understanding the effect that re-work has on work backlogs. In this setting, the customers represent software projects, the servers represent in-house software developers, and the feeding-back flow of customers represents re-work that must be done on projects as the result of problems discovered during independent outside testing or early use. Large delays in discovering problems are common here, and a large-delay limit theorem is thus well suited for this setting. See Bohn's (2000) provocative article, and the references therein, for discussion on how re-work counter-intuitively affects work backlogs in the software development industry.

The organization of the paper is as follows. In Section 2, we give the main theorem and a corollary for Poisson arrival processes and, in Section 3, we prove these results and comment on generalization to the case of nonindependent feedback times.

2. Main results

In the feedback queueing system considered here, customers arrive to the system according to a rate λ stationary ergodic process. The ith customer arrives at the end of the queue and will return an additional M_i times before completely departing from the system, where the M_i are i.i.d. with finite mean m and finite variance σ^2.

After arriving, the ith customer waits in the queue, receives service from one of K servers having a general distribution, and then returns to the back of the queue after a feedback delay. The delay during the kth time feeding back is cX_{ik} time units for a fixed scaling constant $c > 0$. We assume that all the variables X_{ik} are i.i.d. with a common distribution denoted generically by X, with continuous density function f.

We assume for stability that customers arriving to find more than b units of work in the queue immediately skip the queue and are considered lost for that single pass only. Thus, customer i arrives at the back of the queue a total of $1 + M_i$ times, but only actually waits in the queue when the wait is less than b time units. This assumption is of little practical significance if the queue is stable and b is large. No further stability assumptions are needed for the results here.

To construct a stationary version of the feedback queue we first fix t and suppose that the entire system starts empty at time $-t$. Let $A^t_{i,k}$ be the time at which the ith customer arrives at the queue for the kth time (where the customers are ordered by their initial arrival times $A^t_{i,1}$) so that

$$\cdots < A^t_{2,1} < A^t_{1,1} < 0 < A^t_{0,1} < A^t_{-1,1} < A^t_{-2,1} < \cdots .$$

Let

$$\xi^t_k := \sum_{i=-\infty}^{\infty} \delta_{A^t_{i,k}},$$

where δ_x denotes the point mass at x. Our main result concerns properties of the stationary feedback processes

$$\xi_k := \lim_{t\to\infty} \xi^t_k, \qquad k = 1, 2, \ldots,$$

which we assume to exist and be unique. In other words, ξ_k is the stationary process of times when customers arrive at the back of the queue for the kth time. We omit the superscript t on all quantities associated with the queue from now on to indicate quantities for these stationary processes.

We now present the main result of this section, which states that as $c \to \infty$ the process of customers returning to the back of the queue converges to a Poisson process with rate $m\lambda$. Below we let P_x be a Poisson point process with rate x.

Theorem 2.1. *Let $\xi = \sum_{k>1} \xi_k$ be the stationary process of customers returning to the back of the queue after feeding back in the above feedback queueing model having a stationary ergodic rate λ initial arrival process ξ_1. Then*

$$\xi \xrightarrow{\text{D}} P_{m\lambda} \quad \text{as } c \to \infty,$$

where $\xrightarrow{\text{D}}$ denotes convergence in distribution for point processes.

Remark 2.1. Clearly, with a general initial arrival process ξ_1, the superposition $\sum_{k\geq 1} \xi_k$ would not be expected to converge to a Poisson process. Theorem 2.1 only applies to the feeding back customers in $\sum_{k>1} \xi_k$.

We next present a corollary for feedback queues with Poisson arrivals, stating that the effective arrival process of customers—including both the initial arrivals and customers who feed back—converges to a Poisson process.

Corollary 2.1. *Consider the above feedback queue with a Poisson rate λ arrival process. Then with the definitions above*

$$\sum_{k\geq 1} \xi_k \xrightarrow{\text{D}} P_{(m+1)\lambda} \quad \text{as } c \to \infty.$$

Corollary 2.1 motivates the large-delay approximation

$$L_{\text{F}} \overset{\text{D}}{\approx} L_{\text{M/G/}k},$$

meaning that we approximate the stationary queue length, denoted by the random variable L_{F}, using the stationary queue length for an M/G/k queue with no feedback but Poisson rate $(m+1)\lambda$ arrivals, denoted by the random variable $L_{\text{M/G/}k}$. This motivates the use of standard queueing formulas to approximate performance measures for the feedback queue.

Remark 2.2. Breiman (1963) considered the following traffic flow problem. Suppose that traffic on a road at time 0 is distributed according to points in a stationary ergodic point process, and that each car moves at a random velocity (i.i.d. with continuous density). Then, as time approaches infinity, the positions of cars converge in distribution to a Poisson point process. This result follows easily from Theorem 2.1, and our approach gives a different coupling-based proof. Just consider the feedback queue with a single feedback and zero service times. If the velocity distribution is used as the feedback distribution and the process of initial arrivals is viewed as the positions of cars on the road at time 0, then the process of feeding back customers can be viewed as the position of cars at time c. The result then follows.

3. Proof of the main results

Our approach to the proof can be summarized as follows. We first construct another coupled version of the feedback queueing system where there are Poisson arrivals and an infinite number of servers, and we show that here the feedback flow of customers is asymptotically a Poisson process. Here customers operate independently, and thus the feeding-back customer flow will essentially be a superposition of independently translated Poisson processes. We then couple this system to the original system so that asymptotically almost all feeding-back customers in the original system are coupled to feeding-back customers in the infinite-server system, and thus the original system will have an asymptotic Poisson feedback flow. This approach is formalized below.

At first thought it may seem as though we could start a proof by conditioning on the output process of the queue (prior to feeding back) after customers make their first pass through the queue. But, interestingly, it can be seen that, conditional on this, the subsequent feedback times are no longer independent; the second or third pass through the queue for one customer can interfere with the first pass through the queue of a customer who arrives later. With an infinite number of servers, however, there is no customer interference and this type of argument works. Our results here for a finite number of servers do not appear to follow directly from standard point-process convergence results (see Daley and Vere-Jones (1988), for example) in settings with independence, due to the subtle dependencies arising from the interference between customers as they wait in the same queue.

We first define a second feedback queueing system coupled to the original system, which we call the 'infinite-server system'. This system evolves in the same fashion as the original system but has rate λ Poisson arrivals and an infinite number of servers. Quantities associated with this system have the symbol $\tilde{}$ placed above them. Thus, we denote the initial arrival process by $\tilde{\xi}_1 \stackrel{\mathrm{D}}{=} P_\lambda$, and the process of customers who arrive for the kth time by $\tilde{\xi}_k$, with $\tilde{\xi} = \sum_{k>1} \tilde{\xi}_k$. In this model, customers never wait in queue, and thus do not interfere with each other.

Our first lemma states that the feedback process for this infinite-server system is asymptotically a Poisson process. Below we use the notation $\Xi(B)$ to denote the number of points in the set B for the point process Ξ. For some of the theory and notation for point processes, see Daley and Vere-Jones (1988).

Lemma 3.1. *With the above notation,*

$$\tilde{\xi} \stackrel{\mathrm{D}}{\to} P_{m\lambda} \quad as\ c \to \infty.$$

Proof. Since the feedback delay can also be viewed also as an infinite-server queue, we essentially have tandem M/G/∞ queues with 'thinning' between them to account for when customers depart the system. Since the stationary output of an M/G/∞ queue is a Poisson

process, and a 'thinned' Poisson process is still a Poisson process, it follows that

$$\tilde{\xi}_k \stackrel{\mathrm{D}}{=} P_{\lambda_k}, \tag{3.1}$$

where $\lambda_k = \lambda \,\mathrm{P}(M \geq k-1)$. The processes $\tilde{\xi}_k$, $k = 2, 3, \ldots$, are, however, not independent, so their superposition $\tilde{\xi}$ is not necessarily a Poisson process.

Next, fix some bounded Borel set $B \subset \mathcal{R}$ and let $w = \sup\{x : x \in B\} - \inf\{x : x \in B\}$ be the width of the set B. Let N be the number of different customers who return to the back of the queue during B. Note that $\tilde{\xi}(B) \geq N$ since a customer can return more than once to the back of the queue. It can be seen that N has a Poisson distribution with parameter $\int_{-\infty}^{+\infty} \lambda p(x)\,\mathrm{d}x$, where $p(x)$ is the probability that a customer initially arriving at time x ever returns to the back of the queue during B. The chance that some customer returns more than once during B is

$$\mathrm{E}[\mathrm{P}(N > \tilde{\xi}(B) \mid N)] \leq \mathrm{E}[N\,\mathrm{P}(cX \leq w)] \to 0 \quad \text{as } c \to \infty.$$

Thus, with probability approaching 1, all customers returning during B will be different customers, and summing (3.1) over k we see that the rate of the process $\tilde{\xi}$ must be $m\lambda$. We thus have

$$\tilde{\xi}(B) \stackrel{\mathrm{D}}{\to} P_{m\lambda}(B) \quad \text{as } c \to \infty,$$

and the lemma will then follow from this (see Daley and Vere-Jones (1988)).

Our second lemma is a coupling result for random variables with a continuous density function. This type of coupling was used (independently of the present work) by Lindvall (2000) in the context of simulation. We use this later to couple feedback times.

Lemma 3.2. *Given a continuous density function f on $\mathcal{R}^+$ and $\varepsilon > 0$ there exists a $\delta^* > 0$ so that for all δ with $0 < \delta < \delta^*$ it is possible to construct nonnegative random variables X and $\tilde{X}$ on the same probability space so that they both have the same density function f and*

$$\mathrm{P}(\tilde{X} = X + \delta) \geq 1 - \varepsilon.$$

Proof. Given $\varepsilon > 0$, pick L so that $\int_0^L f(x)\,\mathrm{d}x \geq 1 - \frac{1}{2}\varepsilon$. Since the continuous function f restricted to the compact set $[0, L]$ is uniformly continuous, we can find a δ^* so that $0 < \delta < \delta^*$ implies that $|f(x) - f(x-\delta)| \leq \varepsilon/2L$.

Fix any $0 < \delta < \delta^*$ and pick a random point (x, y) uniformly in the area under the graph of the function f. Let $\tilde{X} = x$. If $y \leq f(x-\delta)$, then let $X = x - \delta$. Otherwise, if $y > f(x-\delta)$, pick a random point (x, y) uniformly in the area defined by

$$\{(x, y) : f(x-\delta) \geq y \geq f(x)\},$$

and let $X = x - \delta$.

It can be verified that with this construction

$$\mathrm{P}(X \leq t) = \mathrm{P}(\tilde{X} \leq t) = \int_0^t f(x)\,\mathrm{d}x$$

and

$$\begin{aligned}
\mathrm{P}(\tilde{X} = X + \delta) &\geq \int_0^\infty \min(f(x), f(x-\delta))\,\mathrm{d}x \\
&\geq \int_0^L \left(f(x) - \frac{\varepsilon}{2L}\right)\mathrm{d}x \\
&\geq 1 - \tfrac{1}{2}\varepsilon - \tfrac{1}{2}\varepsilon = 1 - \varepsilon,
\end{aligned}$$

where the second inequality follows by the uniform continuity of f restricted to $[0, L]$.

We also need a third lemma about matching up stationary ergodic point processes. This type of argument was used by Prabhakar *et al.* (1996). We will use this to match up the initial arrival processes.

Lemma 3.3. *Given two jointly stationary ergodic rate λ customer arrival processes A, B and $\varepsilon > 0$ there exists a $\delta > 0$ such that we can pair off customers in A with customers in B so that all pairs arrive within δ time units of each other, and the processes of paired and unpaired customers are both stationary processes, the latter having density less than ε.*

Proof. By the ergodic theorem we can find a sufficiently large δ so that the chance that both $A((0,\delta])$ and $B((0,\delta])$ are between $\delta(\lambda - \frac{1}{2}\varepsilon)$ and $\delta(\lambda + \frac{1}{2}\varepsilon)$ is at least $1-\varepsilon$. Fix this value of δ and let U be a uniform$(0,\delta)$ random variable. Now divide the time axis into intervals of the form $(U + i\delta, U + (i+1)\delta]$ for values of $i \in \mathbb{Z}$. By this construction, in at least the fraction $1-\varepsilon$ of these intervals the number of customers in the A and B processes are within $\varepsilon\delta$ of each other. Pair up customers randomly inside intervals where this occurs, and at most rate ε will be left unpaired. Leave all customers unpaired in intervals were this does not occur, and this will happen to at most rate $\lambda\varepsilon$ customers. Since ε was arbitrary, the result follows.

We are now ready to prove the main theorem.

Proof of Theorem 2.1. The theorem will follow from Lemma 3.1 if we can couple together $\tilde{\xi}$ and ξ so that for any bounded Borel set $B \subset \mathcal{R}$ we have

$$\mathrm{P}(\xi(B) = \tilde{\xi}(B)) \to 1 \quad \text{as } c \to \infty. \tag{3.2}$$

To establish (3.2) we fix $\varepsilon > 0$ and employ Lemma 3.3 to pair up at least the fraction $1-\varepsilon$ of customers in $\tilde{\xi}_1$ with customers in ξ_1 so that they arrive not more than d time units apart. Then pick c sufficiently large so that

$$\frac{d+b+a}{c} < \delta^*,$$

where δ^* is found from applying Lemma 3.2 to the feedback delay density function f, and a is the $1-\varepsilon$ fractile of the service distribution.

We then couple the service times and the number of feedback loops made for the customers in each pair so that they are identical. The stability condition ensures that customers can wait in the queue for at most b time units, so that after any pass through the queue at least the fraction $1-\varepsilon$ of paired customers will depart the queue not more than $a+b+d$ time units apart.

Suppose that for a pair of such customers, the customer in the original system departs the queue $\Delta < a+b+d$ time units after the corresponding customer in the infinite-server system (the same argument will apply if they depart in the opposite order). We couple the pair's corresponding feedback times cX and $c\tilde{X}$ using the approach in Lemma 3.2 with $\delta = \Delta/c < \delta^*$ so that we have

$$\mathrm{P}(c\tilde{X} = cX + \Delta) = \mathrm{P}(\tilde{X} = X + \delta) \geq 1 - \varepsilon.$$

Customers that are initially unpaired are allowed to evolve with uncoupled service and feedback times. This means that with probability at least $1-\varepsilon$ customers which are paired together on arrival at the queue will be coupled to arrive at exactly the same time for the next pass through the queue, and with probability at most ε the pair becomes 'uncoupled'. We make the same coupling for subsequent feedback times, and customers who become uncoupled are allowed to evolve with uncoupled service and feedback times.

Let A be the event that a pair starting out ever gets uncoupled, and let M be the number of times customers in the pair feed back. The previous paragraphs give $\mathrm{P}(A \mid M) \leq 2M\varepsilon$, and thus the expected number of arrivals per customer from customers who eventually become uncoupled equals

$$\mathrm{E}[M\mathbf{1}_A] = \mathrm{E}[M\,\mathrm{P}(A \mid M)] \leq 2\varepsilon\,\mathrm{E}[M^2] = 2\varepsilon(\sigma^2 + m^2).$$

Since this can be made arbitrarily small, the chance of an uncoupled customer arriving during B can also be made arbitrarily small. Since all the remaining customers in B will be paired and coupled, this establishes (3.2) as $\varepsilon \to 0$.

Remark 3.1. The coupling argument above can be easily extended to situations with feedback times which are not i.i.d. Suppose that the kth feedback time for the ith customer cX_{ik} is independent of everything other than k, the total number of passes he will make M_i, and the service times $S_{i1}, \ldots, S_{ik}$ he has experienced so far. Suppose further that X_{ik} has a continuous density which is a bounded function of k, M_i, and $S_{i1}, \ldots, S_{ik}$. It can be easily seen that the same argument goes through to yield the result of Theorem 2.1 for this model as well.

Acknowledgement

We would like to thank an anonymous referee for helpful comments which have greatly improved the presentation of the results in this paper.

References

Barbour, A. D. and Brown, T. (1996). Approximate versions of Melamed's theorem. *J. Appl. Prob.* **33,** 472–489.
Bohn, R. (2000). Stop fighting fires. *Harvard Business Rev.* **78,** No. 4, 82–91.
Breiman, L. (1963). The Poisson tendency in traffic distribution. *Ann. Math. Statist.* **34,** 308–311.
Daley, D. J. and Vere-Jones, D. (1988). An introduction to the theory of point processes. Springer, New York.
D'Avignon, G. R. and Disney, R. L. (1977/78). Queues with instantaneous feedback. *Management Sci.* **24,** 168–180.
Foley, R. D. and Disney, R. L. (1983). Queues with delayed feedback. *Adv. Appl. Prob.* **15,** 162–182.
Jackson, J. R. (1957). Networks of waiting lines. *Operat. Res.* **5,** 518–521.
Kumar, P. R. (1993). Re-entrant lines. *Queueing Systems* **13,** 87–110.
Kuri, J. and Kumar, A. (1997). On the optimal control of arrivals to a single queue with arbitrary feedback delay. *Queueing Systems* **27,** 1–16.
Lindvall, T. (2000). On simulation of stochastically ordered life-length variables. *Prob. Eng. Inf. Sci.* **14,** 1–7.
Melamed, B. (1979). Characterizations of Poisson traffic streams in Jackson queueing networks. *Adv. Appl. Prob.* **11,** 422–438.
Mountford, T. and Prabhakar, B. (1995). On the weak convergence of departures from an infinite series of $\cdot$/M/1 queues. *Ann. Appl. Prob.* **5,** 121–127.
Prabhakar, B., Mountford, T. S. and Bambos, N. (1996). Convergence of departures in tandem networks of $\cdot$/GI/∞ queues. *Prob. Eng. Inf. Sci.* **10,** 487–500.
Takacs, L. (1963). A single-server queue with feedback. *Bell System Tech. J.* **42,** 505–519.
Wortman, M. A., Disney, R. L. and Kiessler, P. C. (1991). The M/GI/1 Bernoulli feedback queue with vacations. *Queueing Systems* **9,** 353–363.